# One-Day Listen-and-Repeat Training of a Non-Native Vowel Duration Contrast for speakers of Namibian Languages

Antti Saloranta[1], Katja Haapanen[1], Kimmo U. Peltola[1], Henna Tamminen[1], Paavo Alku[2], Lannie Uwu-khaeb[3], Maija S. Peltola[1]

1) Phonetics and Learning, Age & Bilingualism laboratory, University of Turku, Turku, Finland
2) Department of Information and Communications Engineering, Aalto University, Espoo, Finland
3) Future Tech Lab, University of Turku in Windhoek, Namibia

**Abstract**

Listen-and-repeat training has previously been successfully used to train the perception and production of non-native vowel quality and duration contrasts. This study used a one-day listen-and-repeat training paradigm for the production of a non-native vowel duration contrast, /tite – ti:te/ with no feedback or other instructions. Learning results were assessed by acoustic analysis of the produced durations, and identification of the productions by listeners with quantity contrasts in their native Finnish language. Training participants were 18 Namibian speakers of various Bantu and Khoe languages. The results showed that the majority of the speakers did not produce a consistent acoustic duration contrast between the target words. In the identification task, the listeners' performance was at essentially chance level for almost all of the speakers. The results are discussed in terms of earlier results using the same stimuli, training design and language background.

*Keywords:* production training; listen-and-repeat training; vowel duration; identification; non-native speech learning

## 1 Introduction

The duration of individual speech segments varies to some extent in virtually all languages, but the significance of this variation is highly language-dependent. Duration can be a cue for or be affected by phenomena such as stress, speech rate and phonetic context (e.g. Cho & McQueen, 2005; Fougeron & Keating, 1997; White & Turk, 2010). In some languages, however, segment duration is a phonological feature, in that it affects the meanings of words. Some of the most well-known of these languages, typically called *quantity languages*, are Finnish, Japanese, Estonian and Hungarian. In Finnish, for example, the duration of both consonants and vowels is a phonologically significant feature, and their duration varies independently of each other and of stress (Suomi, Toivanen, & Ylitalo, 2008, p. 39), such as in the words /tuli/, /tuːli/ and /tulːi/ (“fire”, “wind” and “customs”, respectively).

Native speakers of quantity languages process duration contrasts differently from speakers of other languages. Finnish speakers, for example, are more sensitive to duration differences than speakers of German (Kirmse et al., 2008) and they show a category boundary effect for duration, unlike speakers of Russian (Ylinen, Shestakova, Huotilainen, Alku, & Näätänen, 2006). Learning phonological quantity contrasts can therefore be quite difficult for second language (L2) learners. According to models of second language acquisition, such as the revised Speech Learning Model (SLM-r) (Flege & Bohn, 2021) and the L2 version of the Perceptual Assimilation Model (PAM-L2) (Best & Tyler, 2007), non-native sounds or contrasts that are similar to existing native phonemes, but differ from them in a systematic way, present the most difficult learning situation. In the case of quantity contrasts for a speaker of a non-quantity language, the systematic difference is the duration of the non-native sound. The quality of the speech sound may be entirely familiar, but the learner also has to differentiate between duration

variation that is likely irrelevant in their native language. Speakers of quantity languages, on the other hand, should face much fewer difficulties in perceiving and producing non-native duration contrasts, as the difference already exists in some form in their native language. Indeed, McAllister, Flege and Piske (2002) tested various experienced L2 speakers of Swedish in their perception and production of Swedish vowel quantity contrasts. They showed that speakers of Estonian, a quantity language, fared better at identifying quantity distinctions in Swedish than speakers of English and Spanish, in which vowel duration is not phonologically relevant. Furthermore, the English speakers outperformed the Spanish speakers in both the perception and production of the Swedish quantity contrast. The researchers suggested that this may have been due to phonetic importance of duration in English. This led them to formulate the Feature Hypothesis, which posits that "L2 features not used to signal phonological contrast in L1 will be difficult to perceive for the L2 learner and this difficulty will be reflected in the learner's production of the contrast based on this feature" (McAllister et al., 2002, p. 253).

It should be noted, however, that there is some evidence of non-native vowel duration contrasts being somewhat salient to listeners without such contrasts in their native languages, as suggested, for example, by the Desensitization Hypothesis (Bohn, 1995). He tested the perception of the English /i/ - /ɪ/ continuum that varies in both spectral and durational features, and found that Spanish and Mandarin speakers differentiate the continuum predominantly on duration, despite not having native experience in the use of the duration cue distinctively. English speakers, on the other hand, relied almost exclusively on spectral differences (i.e., quality differences). The Desensitization Hypothesis states that when spectral cues are insufficient for differentiating two vowels due to desensitization caused by previous linguistic experience, learners with no vowel duration contrasts in their native language will use duration as a cue for differentiating them. Kondaurova and Francis (2008) studied L2

English speakers of Russian and Spanish, neither of which have phonological vowel duration contrasts but do differ in their allophonic use of duration to mark phenomena such as stress and following consonant voicing. They examined the two groups' perception of the English /i/ - /ɪ/ continuum and found that both groups classified the continuum based on duration. Their results regarding the allophonic use of duration in native contrasts in the two languages, however, suggested that the use of duration to classify the English continuum could be at least partially explained either by the use of residual L1 learning mechanisms, as per the results by Escudero and Boersma (2004), or by duration being a naturally salient cue even to those listeners with no L1 duration contrasts, as put forth by Bohn (1995).

Laboratory training has been used in a number of studies investigating the perception and production of non-native duration contrasts. The majority of studies have found improvements in the perception of vowel and consonant duration, particularly when using perceptual methods such as identification and discrimination training (Hardison & Okuno, 2022; Hirata, 2004; Hirata, Whitehurst, & Cullings, 2007; Motohashi-Saigo & Hardison, 2009; Okuno, 2014; Okuno & Hardison, 2016; Tajima, Kato, Rothwell, Akahane-Yamada, & Munhall, 2008). The results from Okuno and Hardison (2016) and Hardison and Okuno (2022) are particularly relevant for the current study, as their findings showed improved production of Japanese vowel duration contrasts using only perceptual training. The training participants took part in eight 25-minute sessions of forced-choice identification training over two weeks with or without visual feedback using waveforms depicting minimal pairs of vowel duration contrasts. Production assessment was done by native Japanese listeners (Okuno and Hardison, 2016) or by listeners and acoustic analysis (Hardison & Okuno, 2022). The listeners were instructed to pay attention only to the vowel duration, and then asked to select which word they thought they heard from four options, or "other" if they were unable to decide. The ratings showed an overall improvement in production accuracy in three of the four token types after the training

was complete (Okuno & Hardison, 2016). Furthermore, correlations between acoustic analysis and ratings showed that more native-like vowel durations were associated with higher native-likeness ratings (Hardison & Okuno 2022). Similar results were achieved by Motohashi-Saigo and Hardison (2009) who used audio only and audiovisual identification training in the learning of Japanese geminates for English-speaking beginner students of Japanese. The training consisted of 10 sessions of self-paced training over two weeks with 120 stimulus presentations per session. Perception of geminate consonants improved for both groups, with some effects transferring to novel stimuli and production as well. Production was assessed by native Japanese raters.

In addition to perceptual studies, training studies using production-based listen-and-repeat tasks have also been conducted, both for non-native vowel and consonant quality (e.g. (Akahane-Yamada, McDermott, Adachi, Kawahara, & Pruitt, 1998; Immonen, Alku, & Peltola, 2022; Immonen, Peltola, Tamminen, Alku, & Peltola, 2023; Jähi, Peltola, & Alku, 2015; Peltola, Rautaoja, Alku, & Peltola, 2017; Peltola, Tamminen, Alku, & Peltola, 2015; Saloranta, Tamminen, Alku, & Peltola, 2015; Taimi, Alku, Kujala, Näätänen, & Peltola, 2014; Tamminen & Peltola, 2015; Tamminen, Peltola, Kujala, & Näätänen, 2015) and quantity contrasts (Saloranta, Alku, & Peltola, 2017, 2020; Saloranta, Heikkola, & Peltola, 2022). Saloranta et al. (2015) used a two-day listen-and-repeat training paradigm for the learning of the non-native vowel quality contrast /y/ - /ʉ/ with Finnish participants. The total number of stimulus pair presentations was 150. When enhanced with production instructions, the training was able to help the participants modify their production of the contrast after just one training session of 30 repetitions of the target pair. Listen-and-repeat training has also been used to train non-native duration contrasts. Saloranta et al. (2017) found an improvement in behavioral oddball discrimination of a non-native vowel duration contrast. Using the same training paradigm and contrast, Saloranta et al. (2020) saw improvements in both preattentive perception, indicated

by increased mismatch negativity amplitudes measured by EEG, and behavioral oddball discrimination. The contrast used was /i/ - /iː/, and the paradigm consisted of 160 repetitions of the target contrast in four sessions of training (30 repetitions/contrast) and three recordings (10 repetitions/contrast) over two days with multilingual participants who did not have quantity contrasts in their native languages. The total duration of the training was roughly 30 minutes. These improvements in perception did not transfer to production. However, the participants in both studies by Saloranta et al. seemed able to perceptually differentiate the non-native vowel duration contrast already at the beginning of the experiment, before any training had taken place. They were also able to produce a clear duration difference already at baseline. This may also be the reason for why no production learning took place: the participants may simply have been able to produce the contrast to their satisfaction at the beginning of the experiment, leaving little room for improvement. For consonant duration differences, this type of training has so far proven ineffective even perceptually (Saloranta et al., 2022).

The purpose of the current study was to examine the effects of a one-day listen-and-repeat paradigm for the training of a non-native vowel duration contrast with a previously untested group of speakers of Namibian languages. Previous research has shown that this type of training can cause changes in non-native production in as little as one day, particularly for vowel quality contrasts (Immonen et al., 2022, 2023; Saloranta et al., 2015), and the current study examined the use of this one-day paradigm on duration contrasts that have proven more resistant to training-induced production changes in previous studies than quality contrasts have (e.g. Saloranta et al., 2020). As an important addition to the assessment methods, similarly to Akahane-Yamada et al. (1998) and Okuno and Hardison (2016), changes in the production of the novel contrast were assessed by listeners whose native language contains relevant phonological contrasts, in addition to acoustic analysis of the produced durations. The

use of native listeners in other studies of L2 production is widespread, particularly in the assessment of higher-level features such as accent or intelligibility (e.g. Munro & Derwing, 1995) and their use is recommended in a somewhat recent review of L2 production studies (Thomson & Derwing, 2015). Previous studies on the use of listen-and-repeat for L2 duration have, however, largely relied on acoustic analysis in the assessment of production learning. Using native listeners allows for the detection of other changes the speakers might be making to produce the contrast, such as changing the duration of other sounds in the words to make the target vowel seem longer. In Finnish, for example, the short vowel in the second syllable of a disyllabic CV.CV word is shorter than a corresponding vowel in a CVV.CV word (Suomi et al., 2008, p. 89), increasing the relative duration contrast between the first (stressed) vowel and the second (unstressed) vowel beyond the absolute duration difference and emphasizing the contrast. The two assessment methods in combination were therefore chosen to help assess how well the speakers are able to differentiate the long and short vowels in their productions. It is possible that previous studies, focusing on the acoustic duration of the target vowel, could have missed other changes the speakers were making.

The listen-and-repeat methodology closely resembles phonetic imitation. It has been shown, for example, that L2 English speakers of Polish, in which vowel duration is not a marker of following consonant voicing due to the fact that final obstruent voicing is neutralized, are able to imitate this duration difference from the speech of native English speakers (Zając & Rojczyk, 2014). Studies of the ability to imitate non-native features that do not exist in the imitators native language suggest that pure imitation tasks do not employ the same phonological mechanisms that govern non-imitated speech production, but rather rely on lower level phonetic processing (Hao & de Jong, 2016). Later studies, however, posit that imitation ability is indeed modulated by the representation of the imitated sounds in the imitator's phonological system (e.g. Llompart & Reinisch, 2019; Rojczyk, Sturm, & Przedlacka, 2023). Both

studies found relatively little carry over from imitation to other production tasks, such as reading aloud. This may have been due to phonological representations being stored with lexical information that took over when there was no audio model to follow (Llompart & Reinisch, 2019). Somewhat similar findings have been made by Peltola et al. (2015), who found that misleading orthographic visual stimuli paired with listen-and-repeat directed learners away from correct non-native production. On the other hand, Tamminen (2022) found that learning effects gained with listen-and-repeat training were still retained a year after the original training session. Improvement with listen-and-repeat training could therefore offer indication that the underlying phonological representation have also developed.

The speakers of the current study were all Namibian adults. Namibia is a linguistically diverse country, and while the official language is English, very few people speak it as their native language. The languages most Namibians speak come from the Bantu or Khoe language families. Out of the reported native languages of the speakers in this study, the Khoe language Khoekhoegowab has typically been described as having a five-vowel system /i, e, a, o, u/ with phonologically contrasting duration (Cruttenden, 1992). More recently it has been proposed, however, that this may be a secondary feature in the tone system of the language, where long vowels appear with certain tones and short vowels with others (Fredericks, 2013). Out of the Bantu languages spoken by the participants, Otjiherero, Oshiwambo, Oshikwanyama and Rukwangali are described in literature as having five-vowel systems with short and long vowels (Dammann, 1957; Möhlig, Marten, & Kavari, 2002; Zimmermann & Hasheela, 1998), but long vowels mostly seem to be caused by morphophonological lengthening; in the case of Otjiherero, for example, Möhlig et al. state that “Many apparent long vowels results from phonological processes at morphemic boundaries like consonant elision or vowel coalescence” (Möhlig et al., 2002, p. 17). This is likely to be the case for Subiya and Mbalangwe as well, as they are related to Rukwangali (Van de Velde, Bostoen, Nurse, & Philippson, 2019, p. 44). Two

of the speakers reported English as their first language. The tense-lax contrasts of English use duration as a secondary feature, but the main cue is vowel quality. Out of the non-English L2s reported by the participants, Afrikaans and German also have some contrastive use of vowel duration. In Afrikaans, as in English, there exist tense-lax contrasts differing mainly in vowel quality (Wissing, 2020) and in German there are tense-lax contrasts as well as two phonemic duration contrasts, /ɛ - ɛː/ and /a - aː/ (Altmann, Berger, & Braun, 2012). It therefore seems that most of the speakers have some sort of familiarity with vowel duration differences through the languages they speak, but few use duration as the sole contrasting feature between word meanings, unlike in languages such as Finnish.

Based on the review of the literature and earlier listen-and-repeat studies, the following research questions were formed:

1) Are L1 speakers of Namibian Bantu and Khoe languages able to produce vowel duration differences in unfamiliar words?
2) Can short-term listen-and-repeat training induce changes in their productions?

For question one, we hypothesized that the speakers would be able to produce at least some degree of difference between the short and long vowels (e.g. Bohn, 1995; Kondaurova & Francis, 2008; Saloranta et al., 2017, 2020). Regarding question two, the effectiveness of the training in enhancing this difference is more uncertain. Listen-and-repeat has been shown to be effective over very short periods of training time (Immonen et al., 2022, 2023; Saloranta et al., 2015), but duration contrasts have also shown more resistance to training-induced change than quality contrasts (Saloranta et al., 2020, 2022), at least when measured acoustically. As for the listener identifications, we hypothesized that the listeners would likely focus their attention on the first syllable target vowel in the assessment of the productions, with longer durations resulting in better identifications of long vowels. Due to their native Finnish

language, however, possible changes in the relative durations of the vowels could affect their ratings as well. This could result in identifications being disconnected from the acoustic duration of the target vowel, meaning that there could be more “long” identifications even without a significant increase in the duration of the target vowel. Finally, as the main focus of the study was to examine whether L1 speakers of Namibian languages are able to produce duration contrasts at all, no tests of generalization were included in the paradigm. These could be included in future studies, once the nature of non-native duration processing with speakers of these languages is better understood.

## 2 Materials and methods

### 2.1 Speaker Data

#### 2.1.1 Speakers

All 18 speakers were either students or staff at the University of Namibia. Basic information about the speakers can be found in Table 1. All but two, whose L1 was English, were L1 speakers of Namibian Bantu or Khoe languages. They were all spoken to in English throughout the experiment, and they all volunteered to take part in the study.

**Table 1**

*Basic information about the speakers*

| | |
|---|---|
| **N** | 18 (10 women) |
| **Mean age (range, stdev)** | 23.7 (20–42, 4.7) |
| **First languages (N)** | Khoekhoegowab (5), Oshiwambo (3), Otjiherero (3), English (2), Subia (1), Setswana (1), Oshikwanyama (1), Rukwangali (1), Mbalangwe (1) |
| **All languages spoken (N)** | English (17), Afrikaans (12), Oshiwambo (8), Khoekhoegowab (5), Otjiherero (6), German (2), Subia (1), Setswana (1), Oshikwanyama (1), Mbalangwe (1), Rukwangali (1), Silozi (1), Portuguese (1), Spanish (1), Isiswati (1) |

*Note:* Speaker numbers are based on self-reports by the speakers, who may have accidentally left out some languages they do in fact speak. This is why the number of English speakers is only 17 despite there being 18 participants who all spoke and understood English during the experiment.

### 2.1.2 Stimuli and Training Procedure

The stimuli used in the study were a pair of disyllabic Finnish pseudowords, /tite/ and /tiːte/. Pseudoword is used here to mean that the words follow Finnish phonotactics but are not semantically meaningful. The mean fundamental frequency (F0) was 110 Hz for all vowels, and the duration of the short word was 392 ms, target vowel 154 ms, and the long one 432 ms, target vowel 194 ms. The F1 and F2 for the target vowel were 330 Hz and 2129 Hz, respectively. The duration of the central occlusion was 58 ms. Both members of the stimulus pair were used in the training and recording phases of the task. The stimuli were semisynthetic and based on a male voice. In the semisynthetic method, a glottal signal is extracted from a recording of a real speaker and combined with a digital model of a vocal tract. This results in

stimuli that sound highly natural, but whose phonetic features can be accurately adjusted and controlled. More information on the semisynthetic method can be found in Alku et al. (1999) and on these specific stimuli in Saloranta et al. (2017, 2020). The experiment took place in a small lecture space at the campus of The University of Namibia. The stimuli were presented and recorded with a laptop computer running Sanako Study software and with a Beyerdynamic MMX300 headset.

The one-day training procedure used in the study followed an alternating pattern. First came a baseline recording, followed by a training session, another recording, another training session and a final recording. In both the training and the recording sessions, the stimuli were presented in a short-long alternating pattern. The speakers heard the stimuli and were asked to repeat it as accurately as they could. No other instructions or feedback were provided. In the recordings, the stimulus pair was presented 10 times, and in the training sessions 30 times, for a total of 90 repetitions of the pair. The total duration of the training was approximately 20 minutes.

### 2.2 Identification Task

#### 2.2.1 Listeners

All listeners were university students, who were taking part in basic level phonetics classes. Due to the timeline of speaker data collection, the assessments were collected in three separate sessions. All listeners spoke Finnish as their native language and lived in Finland. In the first session, there were 21 listeners (18 women) aged 20 to 42 (average 23.6). In the second session, there were 11 listeners (9 women) aged 20 to 25 (average 22.5). In the third session there were 13 listeners. Due to a technical problem, gender and age data was lost for this session; however, sessions 2 and 3 were collected a week apart with the students of the same phonetics class, meaning most of the listeners were the same between the two sessions.

Their age and gender are therefore unlikely to be a cause of any major discrepancies in the identifications.

### 2.2.2 Identification tokens

For the identification task, the first and last production of each member of the stimulus pair were extracted from the recordings, i.e. the first production of /tite/ in the first recording, then the first production of /tiːte/, then the final (i.e. $10^{th}$) production of each word from the first recording. This was repeated for each recording session, resulting in 6 versions of each word, for a total of 12 words per speaker. These words served as the tokens in the identification task. The purpose of this pattern of selection was to be able to follow any possible learning effects not only between recording sessions but also within them. In the Results section, the time points within the sessions will be referred to with the format tite/tiite_X_Y, where tite refers to short tokens, tiite to long ones, X is the number of the recording session (1–3) and Y is the number of the within-session stimulus (1 or 2). All identification tokens were normalized to 65 dB SPL in Praat, version 6.2.1.4 (Boersma & Weenink, 2022).

### 2.2.3 Identification procedure

In the identification task, the listeners were asked to evaluate which word they heard the speakers produce. The tokens were presented using an online platform, designed using jsPsych (de Leeuw, 2015). The listeners completed the task on their personal devices, either a laptop computer or a smartphone, using headphones. Before starting the identification, the listeners were able to set the volume of the stimulus presentation to a comfortable level using unrelated speech tokens that had been normalized to the same 65 dB SPL as the identification tokens. Each token was presented once, and the listener was asked to choose which word they heard the speaker produce by pressing an on-screen button. The labels used in the selection were the phonemic transcriptions /tite/ or /tiːte/. They also had the choice of responding

"undecided" if they could not reasonably decide between either option. This could have been the case, for example, if the speaker produced the target vowel with a different vowel quality compared to the stimuli or produced sounds that were not present in the stimuli themselves. The listeners were, however, told to focus mainly on vowel length in their decision-making, in order to limit the use of the "undecided" option.

The tokens were presented one speaker at a time, and each token was presented three times, for a total of 36 tokens per speaker. The presentation order of both the tokens and the speakers was random for each listener. As stated in the previous section, the identifications were collected in three separate sessions, but all followed the same procedure. The numbers of evaluated speakers per session were 5, 7 and 6 for sessions 1–3, resulting in 180, 252 and 216 evaluations per listener, respectively.

**2.3 Acoustic analysis**

All 1080 productions of the stimuli by the speakers were analyzed acoustically. Total duration of each produced word and the duration of the /i/ and /iː/ vowels were measured manually using Praat. No vowels had to be rejected, resulting in 2160 total measurements. These values were then averaged separately for productions of long and short stimuli, resulting in two average values (total word duration and vowel duration) per stimulus for each session, resulting in the final 12 values per speaker used in further analyses. Acoustic analyses of vowel durations were performed using the original recording files with no volume normalization.

In order to reduce differences caused by speech rate and other individual factors, the average vowel duration values obtained from the acoustic analysis for each speaker were normalized. This was done by dividing the average duration of the vowels in the long productions by the vowel durations in the short ones. The resulting ratio is 1.0 when the speaker has produced both the long and short stimuli with equally long vowels. If the ratio is over 1, productions of

the long stimulus have longer vowels, and if it is below 1, the opposite is true. These ratios were used to compare productions of vowel durations in the Results section. The ratio for the long-short stimulus pair used in the study was 1.26. For reference, in Finnish a vowel length contrast in the same position as in the stimuli would typically have a ratio of 2.2–2.4 (Lehtonen, 1970, p. 89; Wiik, 1965, p. 60). This short ratio was chosen in order to avoid ceiling effects, as when listened to in isolation, it is likely that a contrast following typical Finnish ratios would be reliably detected by physical difference alone, regardless of the presence of quantity contrasts in the listeners native language. The chosen contrast was selected from a synthesized continuum of exemplars, varying in the duration of the first syllable vowel, and judged to straddle to the Finnish quantity category boundary by five L1 Finnish speaking phoneticians.

### 2.4 Identification analysis

For analysis, all the responses from the listeners were pooled together for each token in each session, resulting in 63, 33 and 39 assessments for each of them in the first, second and third sessions, respectively. The percentages of /tite/, /tiːte/ and “undecided” responses were calculated. No interlistener agreement assessments were made, and the listeners were essentially considered to be identical for the purposes of the analysis.

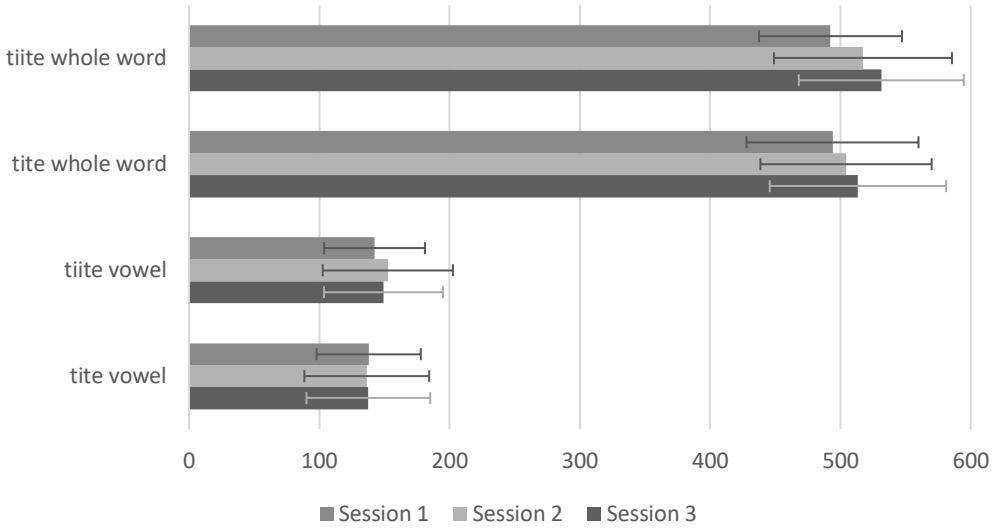
## 3 Results

### 3.1 Acoustic analyses

Mean durations of all productions by all speakers can be seen in Figure 1**Error! Reference source not found.**. Average durations for short and long productions are very similar on the word and segment levels, with the whole words exhibiting slight change towards longer durations with each session. Vowel durations show less change, although some variation can be seen in the long vowels.

**Figure 1.**

*Mean durations of productions, milliseconds*

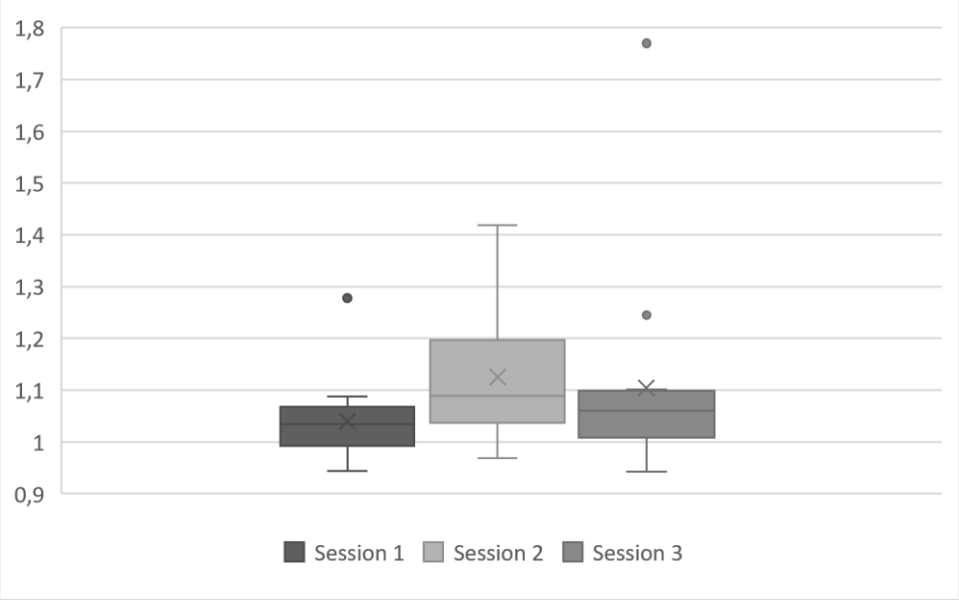



*Note.* Mean durations of whole words and first syllable vowels produced by the whole speaker group. The error bars represent one standard deviation.

The mean long/short ratios of the productions for the entire speaker group can be seen in Figure 2**Error! Reference source not found.**. In the first session, the speakers produced an average ratio of 1.03. In the second session, this increased to 1.13, but fell down to 1.08 in the third. A Session(3) repeated measures Analysis of Variance (ANOVA) was performed for the ratios between the sessions, resulting in a significant main effect of Session ($F$ (2, 34) = 3.816; $p$ = 0,032; $\eta_p^2$ = 0.183). Post hoc analysis with a Bonferroni adjustment showed that there was a significant increase in the mean ratio between Sessions 1 and 2 ($p$ = 0.008) but not between Sessions 1 and 3 or sessions 2 and 3, suggesting the change in production from baseline was temporary. Comparison with the mean durations in Figure 1 shows that the change in the ratios was caused by slight variation in the production of the long vowels.

**Figure 2.**

*Long/short vowel ratios*

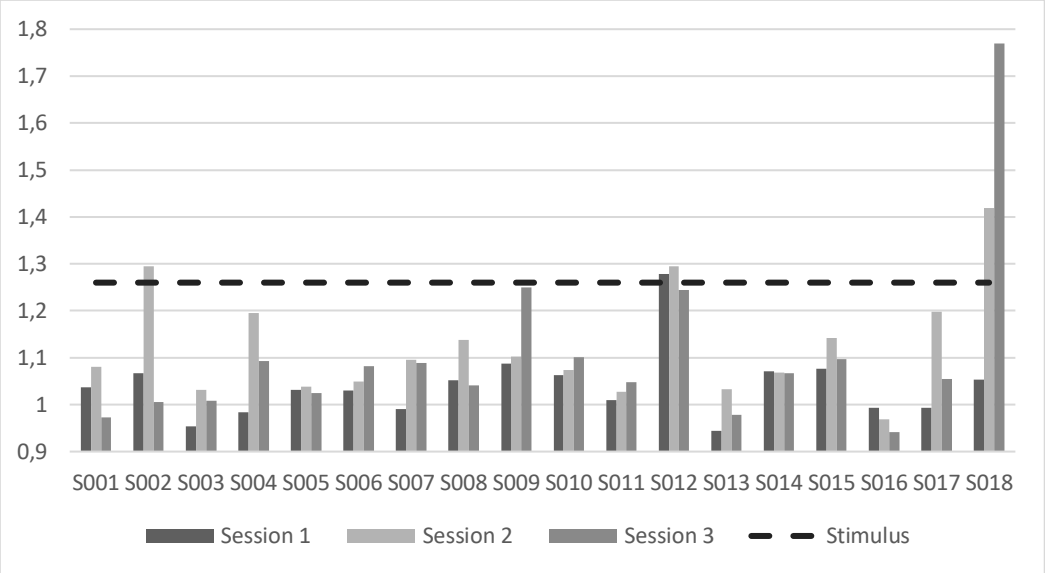



*Note.* Mean long/short ratios of the first syllable vowels produced by the whole speaker group. The mean value is marked by an X.

Examination of individual production ratios (Figure 3) shows that the majority of speakers did not exceed a ratio of 1.1 in any of the recordings, and some did it in one recording but not the others. Ratios of under 1 were also observed for several speakers, indicating that they had repeated the short stimuli with longer vowel durations than the long ones. These differences were minute, however, and any ratios close to 1 on either side should be considered a chance level performance.

**Figure 3.**

*Individual production ratios*




*Note.* Long/short ratios of the mean first syllable vowel durations for each individual speaker.

Some individual speakers did produce ratios more closely resembling those in the stimulus words. Speakers S012 and S018 are especially noteworthy for different reasons: S012 was the only speaker who produced a duration contrast similar to the contrast in the stimuli throughout the experiment, whereas S018 exhibited a steady rise in the ratios from virtually identical in Session 1 to a 77% difference in Session 3. S002 and S009 were also able to reach the stimulus contrast level in individual sessions.

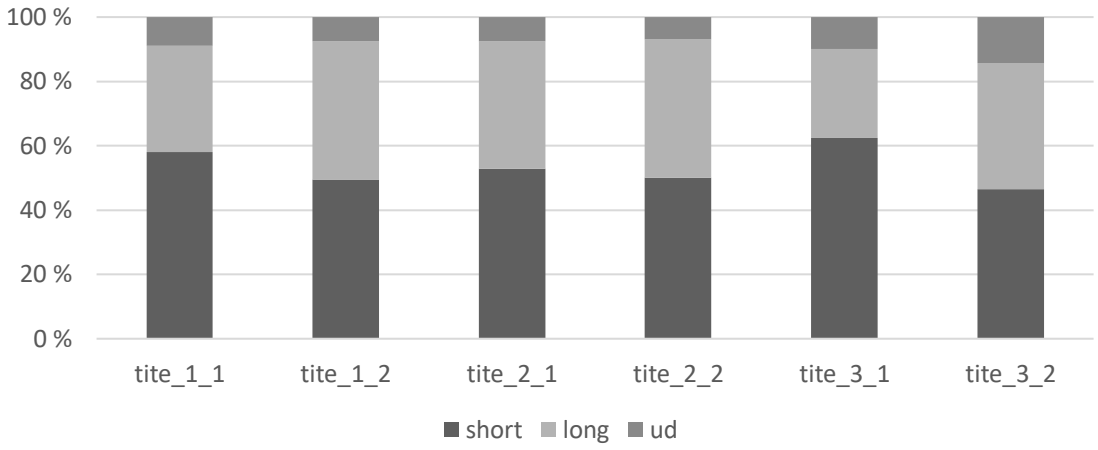
### 3.2 Identifications

The mean identifications for long and short stimuli can be seen in Figure 4 and Figure 5. The short stimuli stayed somewhat constant throughout the experiment, with most productions receiving roughly an equal proportion of short and long classifications. The largest number of short identifications occurs in the first production of the third session, and the lowest in the

final production of the same session. Undecided identifications are mostly under 10%, with the exception of the final stimulus of session 3.

**Figure 4.**

*Mean ratings, short stimuli*

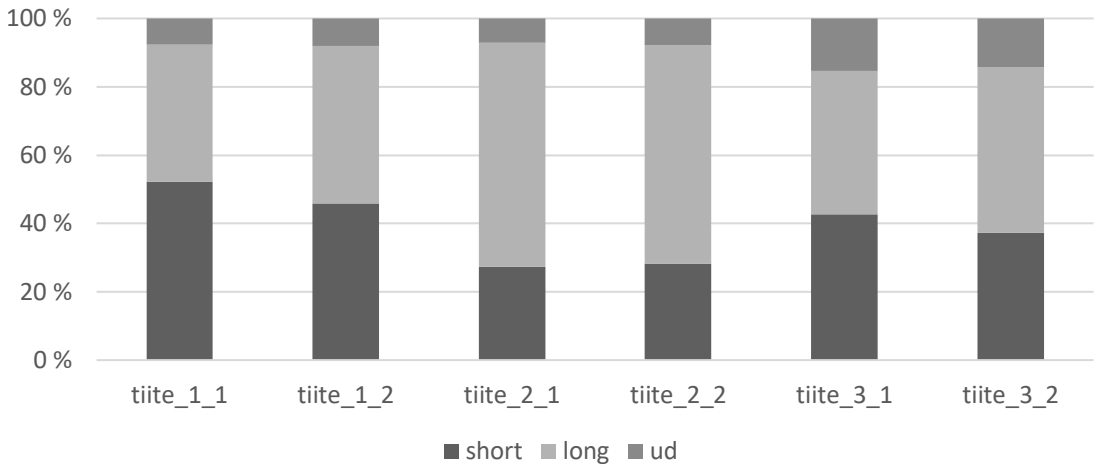



*Note.* Mean proportions of identifications for the short productions at each time point. UD = undecided.

The mean identifications for the productions of the long stimuli show a slightly different pattern, with the second session overall showing the highest proportion of “long” identifications for the long stimuli, with 66% and 64% for the first and last production, respectively. This contrasts with sessions 1 and 3, where the proportions of “long” identifications for the long stimuli were in the range of 40 to 48%. Undecided identifications are under 10% in the first two sessions, and 15% and 14% for the stimuli in the third session.

**Figure 5.**

*Mean ratings, long stimuli*




*Note.* Mean proportions of identifications for the long productions at each time point. UD = undecided.

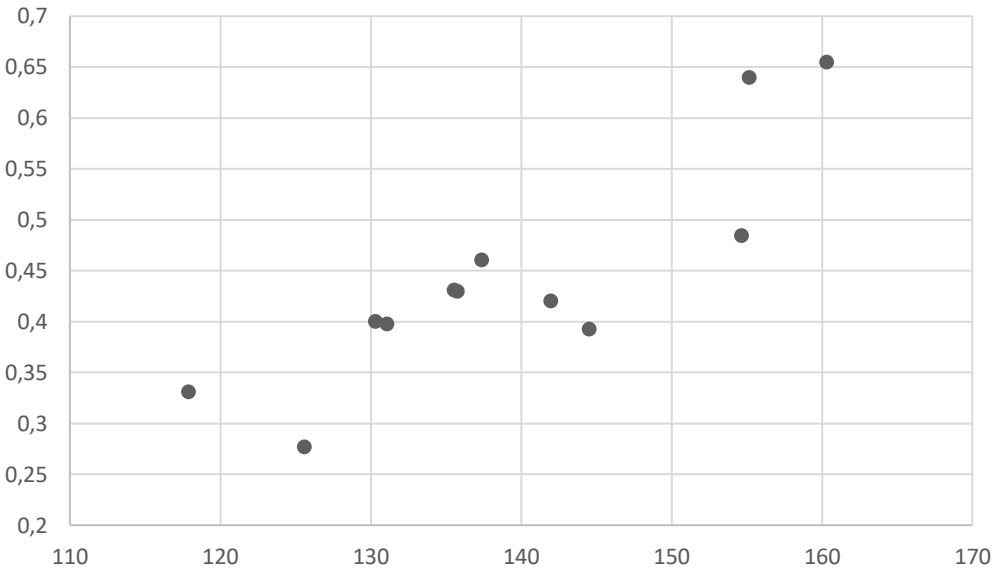
Next, a Stimulus(2) X Session(3) X Time point(2) repeated measures ANOVA was performed with the proportion of "long" identifications in order to examine whether there was a difference between the identifications of the short and long stimuli, whether they changed over time and whether any possible changes showed different patterns between productions of the short and long stimuli. This resulted in a significant main effect of Stimulus ($F$ (1, 17) = 19.506; $p < 0{,}001$; $\eta_p^2 = 0.534$) and a significant Stimulus X Session interaction (F (2, 34) = 3.953; $p < 0{,}029$; $\eta_p^2 = 0.189$), indicating that there was an overall difference in the mean proportion of long identifications between the short and long stimuli, and that the proportion was different between the stimuli in different sessions. The main effect is likely caused by the fact that averaged across all sessions, the mean proportion of "long" identifications for the long stimuli was higher than for the short ones (52% and 37%, respectively). In order to examine the Stimulus X Session interaction further, two-tailed paired samples t-tests were

performed for the “long” identification proportions between long and short stimuli at the same time points, i.e. comparing the proportion of “long” identifications for tite_1_1 vs. tiite_1_1 productions and so on. These revealed statistically significant differences between the “long” identifications of the productions at time points 2_1 ($t(17) = 3.641$; $p = 0.002$; Cohen’s d = 0.82) and 2_2 ($t(17) = 3.547$; $p = 0.002$; Cohen’s d = 0.71), with no other time points reaching significance. The original significance level of 0.05 was Bonferroni corrected to 0.004 to account for the multiple comparisons.

In Figure 6, the mean percentage of “long” identifications vs. the mean absolute durations of the target vowel has been plotted. The somewhat linear rise in the proportion suggests that the listeners were basing their evaluations on the duration of the vowel, rather than some other feature of the productions, such as duration changes in the second, unstressed vowel. The session numbers represented by the dots have purposefully been left out in order to focus solely on the relationship between identifications and duration, rather than correct or incorrect identifications.

**Figure 6**

*Vowel duration and proportion of "long" identifications*

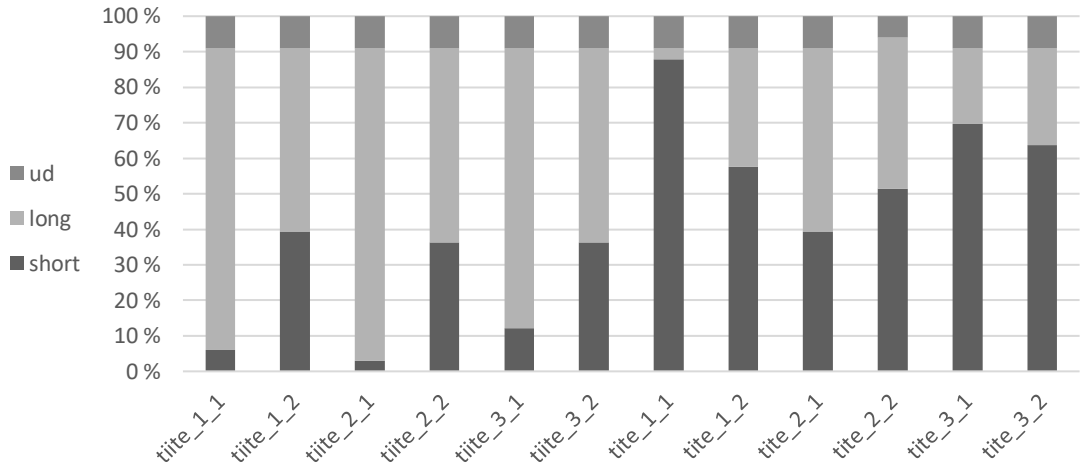



*Note.* Relationship between target vowel duration (X-axis, milliseconds) and the proportion of "long" identifications (Y-axis).

The individual speakers who were best able to produce a difference between the stimulus types in the acoustic analyses received identification results that were generally in agreement with the acoustics. The identification data for speaker S012 (Figure 7), who consistently produced a long/short ratio consistent with the stimuli is presented in Figure 7. It can be seen that their long productions in particular were quite well recognized by the listeners, particularly in the first time point of each session. Data for the short productions are more mixed, but short identifications outnumber long ones in all but one time point, 2_1.

**Figure 7**

*S012 identification proportions*

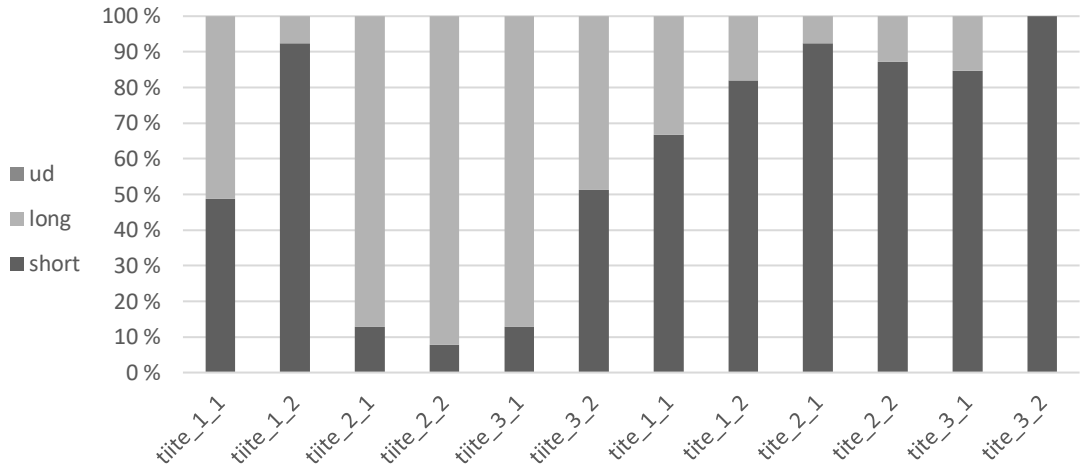



*Note.* Proportions of identifications for speaker S012 for both stimulus types at all time points. UD = undecided.

Speaker S018 (Figure 8) received mainly short identifications for both long and short categories in the first session, but the listeners clearly differentiated between them in the following sessions, with particularly the short productions receiving very few “long” identifications in the final session. In the acoustic analyses of vowel durations, they produced a 5% difference in Session 1, a 42% difference in Session 2, and a 77% difference in Session 3, showing that their differentiation improved over time, consistent with the identification data.

**Figure 8**

*S018 identification proportions*




*Note.* Proportions of identifications for speaker S018 for both stimulus types at all time points. UD = undecided.

### 3.3 Overview of results

Overall, both the acoustic analyses and the listener assessments strongly suggest that most of the speakers did not manage to differentiate the long and short stimuli in their production, either before or after taking part in the training. In the acoustic analyses, this is shown by the similar durations for long and short vowels in most of the speakers' productions. The speakers produced average ratios of 1.03, 1.13 and 1.08 in the first, second and third sessions, respectively. As the ratio of the stimulus pair was 1.26, this means that at best, the speakers were on average able to produce a contrast half the size of what they were hearing. This is especially low given that the difference between the stimulus pair is already quite low. As the total durations of the words remained largely unchanged as well, it is unlikely that the speakers altered any other aspects of their productions in order to better produce the contrast.

The identification data also shows that the speakers did not succeed in consistently producing the stimulus words with distinct durations. When assessed by native Finnish speakers, productions of both the short and long words received significant numbers of both classifications, with 27–52% of the identifications going to the opposite category than what the speaker was repeating. This suggests that the listeners had difficulties consistently deciding what they were hearing.

Examination of the average data for both analyses does show, however, that there was a clear trajectory in the speakers' performances, with Session 2 standing out both in the acoustic analyses and the identifications for the long productions. The difference is confirmed statistically by a significant difference between the first and second sessions in the acoustic analysis, and by a significant difference in the proportion of "long" identifications in both time points of Session 2 versus the other sessions. In the case of the acoustic analyses, the change is towards a larger ratio, and in the identification data it is caused by a larger proportion of "long" identifications for the long stimuli than for the short ones. While both of these changes are seemingly in the right direction, they do not necessarily indicate improved performance. As stated, the acoustic ratio is far below even the difficult contrast of the stimulus pair. As for the identification, while the long productions did receive more "long" identifications on average, they were only correctly categorized 64–66% of the time in the best session. While not a purely chance level performance due to the inclusion of the "undecided" category, it still clearly indicates that the speakers' productions were not readily identifiable by the listeners. Examination of the individual results of the acoustic analysis confirms that with some exceptions, the speakers were not able to reliably produce clear difference between the long and short stimuli in the production task. Notable exceptions were seen in speakers S012 and S018, however, who were able to produce consistent duration differences in individual sessions or throughout the entire experiment.

The overall acoustic results are also supported by the identification data, which shows that higher acoustic ratios are generally paired with more correct categorizations of the tokens, and that longer durations of the target vowel seem to be paired with a higher proportion of "long" evaluations. This suggests that on average, the speaker group did not change anything else about their productions in order to emphasize the duration difference between the short and long vowels. It should be noted, however, that direct comparison between the acoustic analyses and the identification task is impossible, as the acoustic durations are an average of 10 productions, while identification is based on individual productions. The identification tokens may therefore contain mistakes or acoustic phenomena unrelated to the speakers' ability to produce the contrast that may have affected the listeners' decisions.

**4 Discussion and conclusions**

The purpose of this study was to examine the effects of a one-day listen-and-repeat training paradigm on the production of a non-native vowel duration contrast. To this end, two research questions were formed:

1) Are L1 speakers of Namibian Bantu and Khoe languages able to produce vowel duration differences in unfamiliar words?
2) Can short-term listen-and-repeat training induce changes in their productions?

For question one, the results are quite unexpected in terms of earlier research and our own hypotheses, as the results indicated an almost total lack of differentiation of the duration contrast by the speakers at baseline. The Desensitization Hypothesis (Bohn, 1995) posits that listeners with no duration contrasts in their native languages are able to use duration for differentiating vowel contrasts when spectral features are insufficient. Saloranta et al. (2017, 2020) found similar results using the same, spectrally identical stimuli as the ones used in the current study. Despite not having vowel duration contrasts in their native languages, the

participants of those studies produced mean long/short ratios between 1.24 and 1.38 at baseline, pointing at a much higher rate of differentiation than the majority of speakers in the current experiment. The situation is further complicated by the fact that the native languages of most of the current speakers are described as having at least short and long vowels, or in the case of Khoekhoegowab, phonological duration contrasts (Cruttenden, 1992; c.f. Fredericks, 2013). Despite this, only speakers S012 and S018 stand out from the speaker group, and only speaker S018 is a speaker of Khoekhoegowab. The overall poor performance is not clearly explained by the speakers' L2s, either. Most of the participants, including speakers S012 and S018, are L2 speakers of Afrikaans, which has English-like tense-lax contrasts (Wissing, 2020), which suggests that L2 skill in Afrikaans is not inherently helpful. There are also two L2 speakers of German, a language with two phonemic vowel duration contrasts (Altmann et al., 2012) but their performance is completely unremarkable in comparison to the rest of the group. Based on these various pre-existing factors, it would be expected that more than two speakers would show some baseline production ability with the current contrast, but this does not seem to be the case. It may be that typological differences between the native languages of the participants of the studies by Bohn (1995) and Saloranta et al. (2017, 2020) are behind this finding. The vast majority of participants in all of these earlier studies were native speakers of an Indo-European language spoken in Europe, while the current speakers are native speakers of Bantu and Khoe languages only found in Africa. It may be that for these speakers, the duration contrast presented in the current study is too difficult. This is especially likely if the duration difference occurred in a different morphological context than the participants would have expected based on the languages they spoke, i.e. not over a morpheme boundary. The stimuli in the current study, for example, were originally used in the studies by Saloranta et al. (2017, 2020) and they were designed to follow Finnish phonotactics, where a vowel duration contrast in the first syllable is extremely common. This was likely not the case in the

native languages of the current participants, however, resulting in the form of the stimuli seeming more unfamiliar and therefore more difficult to interpret than they were for participants more familiar with European languages. The Feature Hypothesis (McAllister et al., 2002) suggests that an L2 phonological contrast that is not used in the L1 is difficult for an L2 learner to perceive, and this difficulty will also be reflected in production. It is possible that vowel duration was not important enough a feature in the speakers' native languages to enable them to correctly detect it in these particular stimuli, and therefore they were also unable to produce it. Given the uncertain nature of duration in the L1s of the participants, this is certainly a possibility. The standout performances by the two speakers able to produce duration contrasts may stem from individual factors, such as increased childhood exposure to languages containing some use of the duration cue.

Regarding research question two, the effectiveness of the training, the current results both agree and disagree with earlier research. The general lack of significant and lasting production changes is not surprising based on earlier production-based training studies. While listen-and-repeat training has been shown to produce some learning results in vowel quality contrasts in just one day (Immonen et al., 2022, 2023; Saloranta et al., 2015), with vowel duration contrasts it has been less successful (Saloranta et al., 2017, 2020). In these two studies by Saloranta et al., perception of vowel duration contrasts became more accurate with no changes in production, but this may have been caused by the participants' existing ability to differentiate the duration contrast. This discrepancy between perception and production improvements was also seen in a later study by Saloranta et al. (2022), where identification accuracy on vowel duration contrasts further improved from a near-ceiling level after a four-week language course, but production skills remained largely unchanged, or even became worse. Therefore, the current study further confirms that production of duration contrasts may be more resistant to changes with short-term training than perception. Studies where

duration production changes have been achieved (e.g. Okuno, 2014; Okuno & Hardison, 2016) used a more intensive and longer lasting training paradigm than the one used in this study, and it may be that this is required for the production of novel duration contrasts to improve. Further study is warranted in order to find out whether speakers of these languages would more broadly benefit from easier (initial) training conditions, or some other variation in the paradigm that would facilitate their learning of these contrasts.

All in all, the results of the current study confirm that production of vowel duration contrasts is largely resistant to short-term production training. It may be that longer or somehow refined training paradigms are needed to achieve learning results similar to those achieved with vowel quality contrasts. Further studies should, for example, consider using more varied stimuli, longer training times, or changes to the training style itself, such as the incorporation of feedback or adaptive paradigms. More careful adaptation to the participants' native languages may also be warranted.

Acknowledgements

The authors wish to thank Professor Erkki Sutinen and Helvi Haikokolla for their invaluable help throughout the data collection process, all of the volunteer study participants at the University of Namibia, and Sanako Corporation for providing the software used in the training. This work was supported by a grant from Kone Foundation for the project *Dance as a window to endangered languages and the phonetic world*.

**5 References**

Akahane-Yamada, R., McDermott, E., Adachi, T., Kawahara, H., & Pruitt, J. S. (1998). *Computer-based second language production training by using spectrographic representation and HMM-based speech recognition scores*. Retrieved from http://www.isca-speech.org/archive/icslp_1998/i98_0429.html

Alku, P., Tiitinen, H., & Näätänen, R. (1999). A method for generating natural-sounding speech stimuli for cognitive brain research. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *110*(8), 1329–1333. https://doi.org/10.1016/S1388-2457(99)00088-7

Altmann, H., Berger, I., & Braun, B. (2012). Asymmetries in the perception of non-native consonantal and vocalic length contrasts. *Second Language Research*, *28*(4), 387–413. https://doi.org/10.1177/0267658312456544

Best, C. T., & Tyler, M. D. (2007). Non-native and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege* (Vol. 10389, pp. 13–34). Amsterdam: John Benjamins Publishing Company.

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer*. Retrieved from http://www.praat.org/

Bohn, O.-S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 279–304). York Press, Baltimore.

Cho, T., & McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, *33*(2), 121–157. https://doi.org/10.1016/j.wocn.2005.01.001

Cruttenden, A. (1992). Clicks and syllables in the phonology of Dama. *Lingua*, *86*(2), 101–117. https://doi.org/10.1016/0024-3841(92)90031-D

Dammann, E. (1957). *Studien zum Kwangali: Grammatik, Texte, Glossar*. Hamburg: De Gruyter.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(4), 551–585. https://doi.org/10.1017/S0272263104040021

Flege, J. E., & Bohn, O.-S. (2021). The Revised Speech Learning Model (SLM-r). In *Second Language Speech Learning* (pp. 3–83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728–3740. https://doi.org/10.1121/1.418332

Fredericks, N. (2013). *A study of dialectal and inter-linguistic variations of Khoekhoegowab: Towards the determination of the standard orthography* (Doctoral dissertation, University of the Western Cape). University of the Western Cape, Cape Town, South Africa. Retrieved from http://hdl.handle.net/11394/3806

Hao, Y. C., & de Jong, K. (2016). Imitation of second language sounds in relation to L2 perception and production. *Journal of Phonetics*, *54*, 151–168. https://doi.org/10.1016/j.wocn.2015.10.003

Hardison, D. M., & Okuno, T. (2022). L2 Japanese vowel production: A closer look at transfer effects from perception training with waveforms. In S. McCrocklin (Ed.), *Technological Resources for Second Language Pronunciation Learning and Teaching: Research-based Approaches*. Rowman & Littlefield.

Hirata, Y. (2004). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *The Journal of the Acoustical Society of America*, *116*(October 2004), 2384–2394. https://doi.org/10.1121/1.1783351

Hirata, Y., Whitehurst, E., & Cullings, E. (2007). Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates. *The Journal of*

*the Acoustical Society of America*, *121*(6), 3837–3845. https://doi.org/10.1121/1.2734401

Immonen, K., Alku, P., & Peltola, M. S. (2022). Phonetic listen-and-repeat training alters 6–7-year-old children's non-native vowel contrast production after one training session. *Journal of Second Language Pronunciation*, *8*(1), 95–115. https://doi.org/10.1075/jslp.21005.imm

Immonen, K., Peltola, K. U., Tamminen, H., Alku, P., & Peltola, M. S. (2023). Orthography does not hinder non-native production learning in children. *Second Language Research*, *39*(2), 565–577. https://doi.org/10.1177/02676583221076645

Jähi, K., Peltola, M. S., & Alku, P. (2015). Does interest in language learning affect the non-native phoneme production in elderly learners? *Proceedings of the 18th International Congress of Phonetic Sciences*.

Kirmse, U., Ylinen, S., Tervaniemi, M., Vainio, M., Schröger, E., & Jacobsen, T. (2008). Modulation of the mismatch negativity (MMN) to vowel duration changes in native speakers of Finnish and German as a result of language experience. *International Journal of Psychophysiology*, *67*(2), 131–143. https://doi.org/10.1016/j.ijpsycho.2007.10.012

Kondaurova, M. V., & Francis, A. L. (2008). The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *The Journal of the Acoustical Society of America*, *124*(6), 3959–3971. https://doi.org/10.1121/1.2999341

Lehtonen, J. (1970). *Aspects of quantity in standard Finnish* (Doctoral dissertation). University of Jyväskylä, Jyväskylä.

Llompart, M., & Reinisch, E. (2019). Imitation in a Second Language Relies on Phonological Categories but Does Not Reflect the Productive Usage of Difficult Sound Contrasts. *Language and Speech*, *62*(3), 594–622. https://doi.org/10.1177/0023830918803978

McAllister, R., Flege, J. E., & Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *Journal of Phonetics*, *30*(2), 229–258. https://doi.org/10.1006/jpho.2002.0174

Möhlig, W. J. G., Marten, L., & Kavari, J. U. (2002). *A Grammatical Sketch of Herero (Otjiherero)*. Cologne: Rudiger Köppe Verlag.

Motohashi-Saigo, M., & Hardison, D. M. (2009). Acquisition of L2 Japanese Geminates: Training with Waveform Displays. *Language Learning & Technology*, *13*(2), 29–47. https://doi.org/10.1007/s00167-015-3787-1

Munro, M. J., & Derwing, T. M. (1995). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Okuno, T. (2014). Acquisition of L2 Vowel Duration in Japanese by Native English Speakers (Doctoral dissertation, Michigan State U; Vol. 74). Michigan State U. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=mzh&AN=2014302748&site=ehost-live

Okuno, T., & Hardison, D. M. (2016). Perception-production link in L2 Japanese vowel duration: Training with technology. *Language Learning and Technology*, *20*(2), 61–80.

Peltola, K. U., Rautaoja, T., Alku, P., & Peltola, M. S. (2017). Adult Learners and a One-day Production Training – Small Changes but the Native Language Sound System Prevails. *Journal of Language Teaching and Research*, *8*(1), 1–7.

Peltola, K. U., Tamminen, H., Alku, P., & Peltola, M. S. (2015). Non-native production training with an acoustic model and orthographic or transcription cues. *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–5.

Rojczyk, A., Sturm, P., & Przedlacka, J. (2023). Phonetic imitation in L2 speech: Immediate imitation of English consonant glottalization by speakers of Polish. *Language Acquisition*, *0*(0), 1–12. https://doi.org/10.1080/10489223.2023.2253545

Saloranta, A., Alku, P., & Peltola, M. S. (2017). Learning and generalization of vowel duration with production training: Behavioral results. *Linguistica Lettica*, *25*, 67–87.

Saloranta, A., Alku, P., & Peltola, M. S. (2020). Listen-and-repeat training improves perception of second language vowel duration: Evidence from mismatch negativity (MMN) and N1 responses and behavioral discrimination. *International Journal of Psychophysiology*, *147*(November 2019), 72–82. https://doi.org/10.1016/j.ijpsycho.2019.11.005

Saloranta, A., & Heikkola, L. M. (2022). Acquisition of non-native vowel duration contrasts through classroom education: Perception and production affected differently. *Journal of Second Language Pronunciation*. https://doi.org/10.1075/jslp.20040.sal

Saloranta, A., Heikkola, L. M., & Peltola, M. S. (2022). Listen-and-repeat training in the learning of non-native consonant duration contrasts: Influence of consonant type as reflected by MMN and behavioral methods. *Journal of Psycholinguistic Research*. https://doi.org/10.1007/s10936-022-09868-6

Saloranta, A., Tamminen, H., Alku, P., & Peltola, M. S. (2015). Learning of a non-native vowel through instructed production training. *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: University of Glasgow.

Suomi, K., Toivanen, J., & Ylitalo, R. (2008). *Finnish Sound Structure. Phonetics, phonology, phonotactics and prosody*. Oulu: Oulu University Press. Retrieved from http://urn.fi/urn:isbn:9789514289842

Taimi, L., Alku, P., Kujala, T., Näätänen, R., & Peltola, M. S. (2014). The effect of production training on non-native speech sound perception and discrimination in school-aged children: An MMN and behavioural study. *Linguistica Lettica*, *22*, 114–129.

Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., & Munhall, K. G. (2008). Training English listeners to perceive phonemic length contrasts in Japanese. *The Journal of the Acoustical Society of America*, *123*(1), 397–413. https://doi.org/10.1121/1.2804942

Tamminen, H. (2022). *Plasticity in speech perception – effects of learning, age and bilingualism* (Doctoral dissertation, University of Turku). University of Turku, Turku. Retrieved from https://www.utupub.fi/handle/10024/154540

Tamminen, H., & Peltola, M. S. (2015). Non-native memory traces can be further strengthened by short term phonetic training. *Proceedings of the 18th International Congress of Phonetic Sciences*.

Tamminen, H., Peltola, M. S., Kujala, T., & Näätänen, R. (2015). Phonetic training and non-native speech perception—New memory traces evolve in just three days as indexed by the mismatch negativity (MMN) and behavioural measures. *International Journal of Psychophysiology*, *97*(1), 23–29. http://dx.doi.org/10.1016/j.ijpsycho.2015.04.020

Thomson, R. I., & Derwing, T. M. (2015). The Effectiveness of L2 Pronunciation Instruction: A Narrative Review. *Applied Linguistics*, *36*(3), 326–344. https://doi.org/10.1093/applin/amu076

Van de Velde, M., Bostoen, K., Nurse, D., & Philippson, G. (Eds.). (2019). *The Bantu Languages* (2nd ed.). London: Routledge. https://doi.org/10.4324/9781315755946

White, L., & Turk, A. E. (2010). English words on the Procrustean bed: Polysyllabic shortening reconsidered. *Journal of Phonetics*, *38*(3), 459–471. https://doi.org/10.1016/j.wocn.2010.05.002

Wiik, K. (1965). *Finnish and English Vowels*. University of Turku, Turku.

Wissing, D. P. (2020). Afrikaans. *Journal of the International Phonetic Association*, *50*(1), 127–140. https://doi.org/10.1017/S0025100318000269

Ylinen, S., Shestakova, A., Huotilainen, M., Alku, P., & Näätänen, R. (2006). Mismatch negativity (MMN) elicited by changes in phoneme length: A cross-linguistic study. *Brain Research*, *1072*(1), 175–185. http://dx.doi.org/10.1016/j.brainres.2005.12.004

Zając, M., & Rojczyk, A. (2014). Imitation of English vowel duration upon exposure to native and non-native speech. *Poznan Studies in Contemporary Linguistics*, *50*(4), 495–514. https://doi.org/10.1515/psicl-2014-0025

Zimmermann, W., & Hasheela, P. (1998). *Oshikwanyama grammar*. Windhoek, Namibia: Gamsberg Macmillan.

Antti Saloranta (corresponding author)
antti.saloranta@utu.fi
Koskenniemenkatu 4
20014 University of Turku
Finland

Katja Haapanen
katja.haapanen@utu.fi
Koskenniemenkatu 4
20014 University of Turku
Finland

Kimmo U. Peltola
kimmo.peltola@utu.fi
Koskenniemenkatu 4
20014 University of Turku
Finland

Henna Tamminen
henna.tamminen@utu.fi
Koskenniemenkatu 4
20014 University of Turku
Finland

Paavo Alku
paavo.alku@aalto.fi
Aalto University
Department of Information and Communications Engineering
P.O. Box 15600
FI-00076 Aalto, Finland

Lannie Uwu-khaeb
lannie.uwu-khaeb@utu.fi

Maija S. Peltola
maija.peltola@utu.fi
Koskenniemenkatu 4
20014 University of Turku
Finland